REMARKS

In view of the above amendment and the following discussion, the Applicants submit that none of the claims now pending in the application are anticipated or made obvious under the provisions of 35 U.S.C. §§ 102 and 103. Thus, the Applicants believe that all of these claims are now in allowable form.

I. REJECTION OF CLAIMS 1, 4-5, 7, 10-11, 14-15, 17, 20-23, 26-28, 31-32, 35-36 AND 40-42 UNDER 35 U.S.C § 102

Claims 1, 4-5, 7, 10-11, 14-15, 17, 20-23, 26-28, 31-32, 35-36 and 40-42 stand rejected as being anticipated by Nepustil (U.S. Patent No. 6,240,454, issued May 29, 2001, hereinafter "Nepustil"). In response, the Applicants herein amend independent claims 1, 11, 21-23, 32, 41 and 42 to further clarify aspects of the invention that were already claimed previously and respectfully traverse the rejection.

Nepustil teaches a dynamic reconfiguration of network servers. Nepustil discloses a plurality of servers for processing client requests, wherein at least one first server of the plurality of servers has first information and second information related to the first information, for processing portions of the client requests that require the first information and portions of the client requests that require the second information. (See Nepustil, col. 2, ll. 20-46.) If processing on the at least one server becomes excessive, then the at least one server processes the portions of the client requests which require the first information without also processing the portions of the client requests which require the second information. (See *Id.*) The portions of the client request which require the second information is redirected to at least one second server for processing. (See *Id.*)

The Examiner's attention is directed to the fact that Nepustil fails to teach, show or suggest determining a load on a primary server and offloading a processing request to any one of a plurality of offload servers only if a processing threshold is exceeded at the primary server, as positively claimed by the Applicants. Specifically, Applicants' independent claims 1, 11, 21, 22, 23, 32, 41 and 42 recite:

1. A method, in a network comprising a primary server and a plurality of offload servers, for dynamic offloading of processing requests from said primary server to any one of said plurality of offload servers, the method comprising the steps of:

determining a load on said primary server;

if the load on said primary server is less than a first threshold, serving processing requests at said primary server; and

only if the load on said primary server exceeds said first threshold, then offloading at least a portion of said processing requests to any one of said plurality of offload servers. (Emphasis added)

11. A network apparatus comprising a primary server and a plurality of offload servers connected by an IP-based network, for dynamic offloading of processing requests from said primary server to any one of said plurality of offload servers, the network apparatus comprising:

a load controller connected between said network and said primary server;

a memory connected to said load controller and including data and control instructions for operating said primary server to perform the steps of:

determining a load on said primary server;

if said load on said primary server is less than a first threshold, serving processing requests at said primary server; and

only if said load on said primary server exceeds said first threshold, then offloading at least a portion of said processing requests to any one of said plurality of offload servers. (Emphasis added)

21. A system, including an IP network comprising a primary server and a plurality of offload servers, for dynamic offloading of processing requests from said primary server to any one of said plurality of offload servers, the system comprising:

means for determining a load on said primary server;

means for, if said load on said primary server is less than a first threshold, serving the processing requests at said primary server; and

means for, only if said load on said primary server exceeds said first threshold, offloading at least a portion of said processing requests to any one of said plurality of offload servers. (Emphasis added)

22. A program product including control instructions for controlling the operation of a computer, said program product operative with said control instructions to operate said computer in an IP-based network comprising a primary server and a plurality of offload servers to dynamically offload processing requests from said primary server to any one of said plurality of offload servers, the computer operative with said control instructions to perform the steps of:

determining a load on said primary server;

if said load on said primary server is less than a first threshold, serving the

processing requests at said primary server; and

only if said load on said primary server exceeds said first threshold, then offloading at least a portion of said processing requests to any one of said plurality of offload servers. (Emphasis added)

23. A system for allocating processing requirements on a network between a primary server and a plurality of offload servers, comprising:

a load controller connected to said network for receiving processing requests from clients on said network and allocating said processing requests between said primary and offload servers;

a memory connected to said load controller and storing threshold data and control software for analyzing said threshold data and operating said load controller;

said load controller operative with the threshold data and control software to perform the steps of:

periodically evaluating said processing requests to determine a load on said primary server;

if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to any one of said plurality of offload servers; and

only if said load does not exceed said first threshold, directing said processing requests to said primary server. (Emphasis added)

32. A method for allocating processing requirements on an IP network between a primary server and a plurality of offload servers, comprising:

periodically evaluating processing requests to determine a load on said primary server;

if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to any one of said plurality of offload servers; and

only if said processing load does not exceed said first threshold, directing said processing requests to said primary server. (Emphasis added)

41. A system for allocating processing requirements on an IP network between a primary server and a plurality of offload servers, comprising:

means for periodically evaluating processing requests to determine a load on said primary server;

means for, if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to any one of said plurality of offload servers; and

means for, only if said processing load does not exceed said first threshold, directing said processing requests to said primary server. (Emphasis added)

42. A program product including control instructions for controlling the operation of a computer, said program product operative with said control instructions to operate said computer in an IP-based network comprising a primary server and a plurality of offload servers to dynamically offload processing requests from said primary server to any one of said plurality of offload servers, the computer operative with said control instructions to perform the steps of:

periodically evaluating processing requests to determine a load on said primary server;

if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to any one of said plurality of offload servers; and

only if said load does not exceed said first threshold, directing said processing requests to said primary server. (Emphasis added)

Applicants' invention is directed to a system and method for dynamically allocating processing on a network amongst multiple network servers. As the Internet continues to grow, so does traffic (*e.g.*, processing requests) directed to popular Internet servers on the World Wide Web. To support these high traffic rates, many techniques have been developed that use offload servers associated with primary web servers to process some of the traffic targeted to primary web servers. In many conventional cases, traffic is directed to these offload servers regardless of the current resource availability at the primary web servers. However, it is normally only during peak traffic periods that the offload servers are actually needed; during non-peak periods, substantial amounts of processing resources at the primary server may go unused due to the use of the offload servers. Thus, resources are wasted by using offload servers when they are not needed.

Applicants' invention provides a method for dynamically offloading traffic from a primary server to any one of a plurality of offload servers based on the current load at the primary server and one or more thresholds. In one particular embodiment, the load at the primary server is first determined. If the load falls below a first threshold (*e.g.*, the primary server is not currently over-loaded), then the traffic is directed to the primary server. However, if the load at the primary server exceeds the first threshold (*e.g.*, the primary server is over-loaded or is currently processing the maximum desirable amount of traffic), then at least a portion of the traffic is directed to any one of the plurality of

offload servers. The first threshold may be based on a number of parameters, including network load, CPU utilization, connections per second, various bandwidth loads, various memory loads and the like. In this way, a web site can make use of excess processing capacity, offloading traffic only when the traffic exceeds the web site's processing capacity. Notably, the re-directed traffic may be offloaded to any offload server. Generally, the decision as to which offload server to re-direct traffic to is based on a number of factors such as, client identity, price structure and offload server availability. (See Applicants' specification, pg. 11, l.28 – pg. 13, l. 4.) Thus, the Applicants' invention provides cost savings by drastically reducing offloaded work.

In contrast, Nepustil teaches away from Applicants' invention because Nepustil specifically teaches that only specific portions of client requests are re-directed and they are re-directed to specific offload servers. In other words, each specific offload server is only capable of handling a portion of a client's request. Furthermore, since each offload server only carries a specific portion of the duplicated data, only specific offload server is able to handle specific offloaded requests. (See Nepustil, col. 3, ll. 21-39.) To illustrate, Nepustil specifically teaches:

> "server 105 is a supplemental server for a portion of BZ of server's 106 database B; server 106 is a supplemental server for a portion AY of server's 105 database A and a portion of CW of server's 107 database C; and server 107 is a supplemental server for a portion AX of server's 105 database A." (See *Id.*)

Nepustil further illustrates specific assignments of receiving re-directed client requests in FIG. 2. Unlike Nepustil's teaching of specific assignments, the Applicants' invention teaches that processing requests, only if necessary, may be offloaded to any one of a plurality of offload servers.

Furthermore, the Applicants' invention teaches a primary server and a plurality of offload servers. In other words, these are two distinct type of independent servers where the offload servers may be provided by one or more offload service providers. (See Applicants' specification, pg. 5, ll. 6-11.) In contrast, Nepustil teaches that each server is both a primary server and an offload server. (See Nepustil, col. 3, ll. 30-34.) Thus, in Nepustil if each primary server is above a load limit, then each of the primary

server will have no capacity left to serve as an offload server, i.e., there would be no offload server to re-direct traffic to. Thus, Nepustil teaches away from Applicants' invention. In the Applicants' invention, offload servers are not also primary servers, therefore, an offload server will always be available to receive offloaded process requests. Consequently, Nepustil clearly fails to teach the Applicants' invention and to anticipate independent claims 1, 11, 21, 22, 23, 32, 41 and 42.

Furthermore, dependent claims 4-5, 7, 10, 14-15, 17, 20, 26-28, 31, 35-36 and 40 depend, either directly or indirectly, from claims 1, 11, 21, 22, 23, and 32, and recite additional limitations. As such, and for at least the exact same reason set forth above, the Applicants submit that claims 4-5, 7, 10, 14-15, 17, 20, 26-28, 31, 35-36 and 40 are also patentable and not anticipated by Nepustil. As such, the Applicants respectfully request the rejection of claims 1, 4-5, 7, 10-11, 14-15, 17, 20-23, 26-28, 31-32, 35-36 and 40-42 under 35 U.S.C. § 102 be withdrawn.

II. REJECTION OF CLAIMS 2-3, 6, 8-9, 12-13, 16, 18-19, 24-25, 29-30, 33-34 AND 37-39 UNDER 35 U.S.C. § 103

1. Claims 2-3, 6, 12-13, 16, 24-25, 29, 33-34 and 37-39

Claims 2-3, 6, 12-13, 16, 24-25, 29, 33-34 and 37-39 stand rejected as being obvious over the Nepustil in view of Swildens et al. (U.S. Patent No. 6,694,358, issued February 17, 2004, hereinafter "Swildens"). The Applicants respectfully traverse the rejection.

The teachings of Nepustil are discussed above. Swildens teaches a performance computer network method. Specifically, Swildens teaches a load balancing method that determines the traffic loads (*e.g.*, volume of processing requests) on a plurality of web servers. These various traffic loads are then compared to identify the web server that has the smallest traffic load among the plurality of web servers, and traffic is directed to this server.

The Examiner's attention is directed to the fact that Nepustil and Swildens, singly and in any permissible combination, fail to teach, show or suggest determining a load on a primary server and offloading a processing request to any one of a plurality of

offload servers only if a processing threshold is exceeded at the primary server, as positively claimed by the Applicants' independent claims 1, 11, 21, 22, 23, 32, 41 and 42. (See *supra*.) Applicants' invention is directed to a system and method for dynamically allocating processing on a network amongst multiple network servers. Applicants' invention provides a method for dynamically offloading traffic from a primary server to any one of the plurality of offload servers based on the current load at the primary server and one or more thresholds. In one particular embodiment, the load at the primary server is first determined. If the load falls below a first threshold (*e.g.*, the primary server is not currently over-loaded), then the traffic is directed to the primary server. However, if the load at the primary server exceeds the first threshold (*e.g.*, the primary server is over-loaded or is currently processing the maximum desirable amount of traffic), then at least a portion of the traffic is directed to any one of the at least one offload server. The first threshold may be based on a number of parameters, including network load, CPU utilization, connections per second, various bandwidth loads, various memory loads and the like. In this way, a web site can make use of excess processing capacity, offloading traffic only when the traffic exceeds the web site's processing capacity. Notably, the re-directed traffic may be offloaded to any of the plurality of offload servers. (See Applicants' specification, pg. 11, l.28 – pg. 13, l. 4.)

As discussed above, the alleged combination (as taught by Nepustil) fails to teach, show or suggest the Applicants' invention. Nepustil teaches away from Applicants' invention because Nepustil specifically teaches that only specific portions of client requests are re-directed and they are re-directed to specific offload servers. In other words, each specific offload server is only capable of handling a portion of a client's request. Furthermore, since each offload server only carries a specific portion of the duplicated data, only specific offload server is able to handle specific offloaded requests. (See Nepustil, col. 3, ll. 21-39.) Nepustil further illustrates specific assignments of receiving re-directed client requests in FIG. 2. Unlike Nepustil's teaching of specific assignments, the Applicants' invention teaches that processing requests, only if necessary, may be offloaded to any one of a plurality of offload servers.

The Examiner then asserts that the substantial gap left by Nepustil is bridged by

Swildens. The Applicants respectfully submit that Nepustil and Swildens cannot be meaningfully combined because Swildens teaches away from Nepustil. Nepustil teaches specific offload server assignment, as discussed above. In contrast, Swildens teaches that an optimal sever is chosen based on various factors. (See Swildens, col. 2, ll. 46-64.)

Furthermore, Swildens fails to teach or to suggest a preference for the use of a specified or primary server (*e.g.*, to reduce offloading costs). Rather Swildens requires the monitoring of a plurality of servers that may potentially be used to process traffic. Thus, Swildens clearly teaches away from the Applicants' invention because Swildens requires more data collection than the Applicants' invention does. Applicants' invention determines the load on only one server (e.g., the primary server), unlike Swildens, which determines the load on a plurality of servers. Because it monitors fewer servers, the Applicants' invention requires less calculations and less comparisons and is less time consuming than determining the load traffic data at a plurality of servers, comparing the loads of each server to the loads of other servers and/or respective thresholds and determining which web server to use for processing from among the plurality.

Therefore, Applicants respectfully submit that independent claims 1, 11, 21, 22, 23, 32, 41 and 42 are clearly patentable and not made obvious by Nepustil in view of Swildens.

Furthermore, dependent claims 2-3, 6, 12-13, 16, 24-25, 29, 33-34 and 37-39 depend, either directly or indirectly, from claims 1, 11, 21, 22, 23, and 32 and recite additional limitations. As such, and for at least the exact same reason set forth above, the Applicants submit that claims 2-3, 6, 12-13, 16, 24-25, 29, 33-34 and 37-39 are also patentable and not made obvious by Nepustil in view of Swildens. As such, the Applicants respectfully request the rejection of claims 2-3, 6, 12-13, 16, 24-25, 29, 33-34 and 37-39 under 35 U.S.C. § 103 be withdrawn.

2. Claims 8-9, 18-19 and 30

Claims 8-9, 18-19 and 30 stand rejected as being obvious over Nepustil and

Swildens in view of the Gupta et al. patent (U.S. Patent No. 6,374,305, issued April 16, 2002, hereinafter "Gupta"). The Applicants respectfully traverse the rejection.

The teachings of Nepustil and Swildens have been discussed above. Gupta teaches a web applications interface system in a mobile-based client-server system. Specifically, Gupta teaches architecture that incorporates two specialized software layers: a specialized "proxy" layer that resides on a mobile client station, and a "web agent" that resides on a server. These layers employ respective memory caches and intelligent filtering capabilities, thereby reducing redundant or otherwise unwanted message transmission.

As discussed above, Nepustil and Swildens fail to teach, show or suggest determining a load on a primary server and offloading a processing request to any one of a plurality of offload servers only if a processing threshold is exceeded at the primary server, as positively claimed by the Applicants' independent claims 1, 11 and 23. Gupta fails to bridge this gap in the teachings of Nepustil and Swildens. Thus, claims 1, 11 and 23 are not made obvious by Nepustil and Swildens in view of Gupta.

Dependent claims 8-9, 18-19 and 30 depend, either directly or indirectly, from claims 1, 11 and 23 and recite additional limitations. As such, and for at least the exact same reasons set forth above, the Applicants submit that claims 8-9, 18-19 and 30 are also not made obvious by the teachings of Nepustil and Swildens in view of Gupta.

III. CHANGE OF CORRESPONDENCE ADDRESS

The Applicants again resubmit an Authorization to Act in a Representative Capacity that was previously filed on May 4, 2005. It should be noted that the Applicants have also requested a change of Correspondence Address in the Authorization to Act in a Representative Capacity. It is respectfully requested that the USPTO acknowledges this change.

IV. CONCLUSION

Thus, the Applicants submit that all of the presented claims fully satisfy the requirements of 35 U.S.C. §§102 and 103. Consequently, the Applicants believe that all

of the presented claims are presently in condition for allowance. Accordingly, both reconsideration of this application and its swift passage to issue are earnestly solicited.

If, however, the Examiner believes that there are any unresolved issues requiring the maintenance of the present final action in any of the claims now pending in the application, it is requested that the Examiner telephone Mr. Kin-Wah Tong, Esq. at (732) 530-9404 so that appropriate arrangements can be made for resolving such issues as expeditiously as possible.

Respectfully submitted,

June 5, 2006
Date

Kin-Wah Tong, Attorney
Reg. No. 39,400
(732) 530-9404

Patterson & Sheridan, LLP
595 Shrewsbury Avenue
Shrewsbury, New Jersey 07702